MECHANISM THEORY

Matthew O. Jackson
Stanford University, Stanford, California, 94305, USA.

Keywords: mechanism, mechanism design, dominant strategy, public goods, auction, bargaining, Bayesian equilibrium, Bayesian incentive compatibility, revelation principle, efficiency, individual rationality, balance, strategy-proof, direct mechanism, social choice function, single-peaked preferences, implementation.

Contents

Summary

Some of the basic results and insights of the literature on mechanism design are presented. In that literature game theoretic reasoning is used to model social institutions as varied as voting systems, auctions, bargaining protocols, and methods for deciding on public projects. A theme that comes out of the literature is the difficulty of finding mechanisms compatible with individual incentives that simultaneously result in efficient decisions (maximizing total welfare), the voluntary participation of the individuals, and balanced transfers (taxes and subsidies that net to zero across individuals). This is

explored in the context of various incentive compatibility requirements, public and private goods settings, small and large societies and forms of private information held by individuals.

1. Introduction

The design of the institutions through which individuals interact can have a profound impact on the results of that interaction. For instance, whether an auction is conducted with sealed bids versus oral ascending bids can have an impact on what bidders learn about each other's valuations and ultimately how they bid. Different methods of queuing jobs and charging users for computer time can affect which jobs they submit and when they are submitted. The way in which costs of a public project are spread across a society can affect the decision of whether or not the project is undertaken.

The theory of mechanism design takes a systematic look at the design of institutions and how these affect the outcomes of interactions. The main focus of mechanism design is on the design of institutions that satisfy certain objectives, assuming that the individuals interacting through the institution will act strategically and may hold private information that is relevant to the decision at hand. In bargaining between a buyer and a seller, the seller would like to act as if the item is very costly thus raising the price, and the buyer would like to pretend to have a low value for the object to keep the price down. One question is whether one can design a mechanism through which the bargaining occurs (in this application, a bargaining protocol) to induce efficient trade of the good - so that successful trade occurs whenever the buyer's valuation exceeds that of the seller. Another question is whether there exists such a mechanism so that the buyer and seller voluntarily participate in the mechanism.

The mechanism design literature, models the interaction of the individuals using game theoretic tools, where the institutions governing interaction are modeled as mechanisms. In a mechanism each individual has a message (or strategy) space and decisions result as a function of the messages chosen. For instance, in an auction setting the message space would be the possible bids that can be submitted and the outcome function would specify who gets the object and how much each bidder pays as a function of the bids submitted. Different sorts of assumptions can be examined concerning how individuals choose messages as a function of their private information, and the analysis can be applied to a wide variety of contexts. The analysis also allows for transfers to be made among the individuals, so that some might be taxed and others subsidized (as a function of their private information) to bring their incentives into alignment.

A theme that comes out of the literature is that it in many settings it is impossible to find mechanisms compatible with individual incentives that simultaneously result in efficient decisions (maximizing total welfare), the voluntary participation of the individuals, and balanced transfers (taxes and subsidies that always net out across individuals). Nevertheless, there are some important settings where incentives and efficiency are compatible and in other settings a "second-best" analysis is still possible. This is described in detail in what follows, in the context of different incentive compatibility requirements, public and private goods settings, small and large societies, and forms of private information held by individuals.

2. A General Mechanism Design Setting

Individuals

A finite group of individuals interact. This set is denoted $N = \{1, 2, ..., n\}$ and generic individuals are represented as i, j, and k.

Decisions

The set of potential social decisions is denoted D, and generic elements are represented as d and d'.

The set of decisions may be finite or infinite depending on the application.

Preferences and Information

Individuals hold private information. Individual i's information is represented by a type θ_i which lies in a set Θ_i. Let $\theta = (\theta_i, ... , \theta_n)$ and $\Theta = \times_i\Theta_i$.

Individuals have preferences over decisions that are represented by a utility function v_i: $D \times \Theta_i \to \mathbb{R}$. So, $v_i(d,\theta_i)$ denotes the benefit that individual i of type $\theta_i \in \Theta_i$ receives from a decision $d \in D$. Thus, $v_i(d,\theta_i) > v_i(d',\theta_i)$ indicates that i of type θ_i prefers decision d to decision d'.

The fact that θ_i 's preferences depend only on θ_i is commonly referred to as being a case of *private values*. In private values settings θ_i represents information about i's preferences over the decisions. More general situations are discussed in Section 4.7 below.

Example 1 A Public Project

A society is deciding on whether or not to build a public project at a cost c. For example, the project might be a public swimming pool, a public library, a park, a defense system, or any of many public goods. The cost of the public project is to be divided equally. Here $D = \{0, 1\}$ with 0 representing not building the project and 1 representing building the project.

The value of each individual i from use of the public project is represented by θ_i. In this case, the net benefit of i is 0 from not having a project built and $\theta_i - \frac{c}{n}$ from having a project built. The utility function of i can then be represented as

$$v_i(d,\theta_i) = d\theta_i - d\frac{c}{n}. \tag{1}$$

Example 2 A Continuous Public Good Setting

In Example 1 the public project could only take two values: being built or not. There was no question about its scale. It could be that the decision is to choose a scale of a public project, such as how large to make a park, and also to choose an allocation of the costs. Let $y \in \mathbb{R}_+$ denote the scale of the public project and $c(y)$ denote the total cost of the project as it depends on the scale. Here $D = \{(y, z_1, ..., z_n) \in \mathbb{R}_+ \times \mathbb{R}^n \,|\, \sum_i z_i = c(y)\}$.

Example 3 Allocating a Private Good

An indivisible good is to be allocated to one member of society. For instance, the rights to an exclusive license are to be allocated or an enterprise is to be privatized. Here, $D = \{d \in \{0, 1\}^n : \sum_i d_i = 1\}$, where $d_i = 1$ denotes that individual i gets the object. If individual i is allocated the object, then i will benefit by an amount θ_i, so $v_i(d,\theta_i) = d_i\theta_i$.

Clearly, there are many other examples that can be accommodated in the mechanism design analysis as the formulation of the space D has no restrictions.

Decision Rules and Efficient Decisions

It is clear from the above examples that the decision a society would like to make will depend on the θ_i's. For instance, a public project should only be built if the total value it generates exceeds its cost. A decision rule is a mapping $d : \Theta \to D$, indicating a choice $d(\theta) \in D$ as a function of θ. A decision rule $d(\cdot)$ is efficient if

$$\sum_i v_i(d(\theta),\theta_i) \geq \sum_i v_i(d', \theta_i) \qquad (2)$$

for all θ and $d' \in D$.

This notion of efficiency takes an ex-post perspective. That is, it looks at comparisons given that that the θ's are already realized, and so may ignore improvements that are obtainable due to risk sharing in applications where the d's may involve some randomization. This notion of efficiency looks at maximization of total value and then coincides with Pareto efficiency only when utility is transferable across individuals. Transferability is the case treated in most of the literature.

In the public project example (Example 1), the decision rule where $d(\theta) = 1$ when $\sum_i \theta_i > c$ and $d(\theta) = 0$ when $\sum_i \theta_i < c$ (and any choice at equality) is efficient.

Transfer Functions

In order to provide the incentives necessary to make efficient choices, it may be necessary to tax or subsidize various individuals. To see the role of such transfers, consider the example of the public project above. Any individual i for whom $\theta_i < \frac{c}{n}$ would rather not see the project built and any individual for whom $\theta_i > \frac{c}{n}$ would rather not see the project built. Imagine that the government simply decides to poll individuals to ask for their θ_i and then builds the project if the sum of the announced θ_i 's is greater than *c*. This would result in an efficient decision if the θ_i's were announced truthfully. However, individuals with $\theta_i < \frac{c}{n}$ have an incentive to underreport their values and say they see no value in a project, and individuals for whom $\theta_i > \frac{c}{n}$ have an incentive to overreport and say that they have a very high value from the project. This could result in a wrong decision. (Similarly, if the decision is simply made by a majority vote of the population, the number who vote yes will simply be the number for whom $\theta_i > \frac{c}{n}$. This can easily result in not building the project when it is socially efficient, or building it when it is not socially efficient.) To get a truthful revelation of the θ_i's, some adjustments need to be made so that individuals are taxed or subsidized based on the announced θ_i's and individuals announcing high θ_i's expect to pay more.

Adjustments are made by a transfer function $t : \Theta \to \mathbb{R}^n$. The function $t_i(\theta)$ represents the payment that *i* receives (or makes if it is negative) based on the announcement of types θ.

Social Choice Functions

A pair *d, t* will be referred to as a social choice function, and at times denoted by *f*. So, $f(\theta) = (d(\theta), t(\theta))$.

The utility that *i* receives if $\hat{\theta}$ is the "announced" vector of types (that operated on by $f = (d, t)$) and *i*'s true type is θ_i is

$$u_i(\hat{\theta}, \theta_i, d, t) = v_i(d(\hat{\theta}), \theta_i) + t_i(\hat{\theta}). \tag{3}$$

This formulation of preferences is said to be quasi-linear.

Feasibility and Balance

A transfer function *t* is said to be feasible if $0 \geq \sum_i t_i(\theta)$ for all θ.

If *t* is not feasible then it must be that transfers are made into the society from some outside source. If the *t* is feasible, but results in a sum less than zero in some circumstances, then it generates a surplus which would either have to be wasted or returned to some outsider. (It is important that the surplus not be returned to the society. If it were returned to the society, then it would result in a different transfer function and different incentives.)

A transfer function *t* is balanced if $\sum_i t_i(\theta) = 0$ for all θ.

Balance is an important property if we wish the full (*d, t*) pair to be efficient rather than just *d*. If $\sum_i t_i < 0$, then there is some net loss in utility to society relative to an efficient decision with no transfers.

Mechanisms

A mechanism is a pair *M*, *g*, where $M = M_1 \times \cdots \times M_n$ is a cross product of message or strategy spaces and $g : M \to D \times \mathbb{R}^n$ is an outcome function. Thus, for each profile of messages $m = (m_1, ..., m_n)$, $g(m) = (g_d(m), g_{t,1}(m), ..., g_{t,n}(m))$ represents the resulting decision and transfers.

A mechanism is often also referred to as a game form. The terminology game form distinguishes it from a game (see game theory), as the consequence of a profile of messages is an outcome rather than a vector of utility payoffs. Once the preferences of the individuals are specified, then a game form or mechanism induces a game. Since in the mechanism design setting the preferences of individuals vary, this distinction between mechanisms and games is critical.

3. Dominant Strategy Mechanism Design

The mechanism design problem is to design a mechanism so that when individuals interact through the mechanism, they have incentives to choose messages as a function of their private information that leads to socially desired outcomes. In order to make predictions of how individuals will choose messages as a function of their private information, game theoretic reasoning is used (see *Game Theory*). We start, as much of the literature on mechanism design did, by looking at the notion of dominant strategies, which identifies situations in which individuals have unambiguously best strategies (messages).

3.11 Dominant Strategies

A strategy $m_i \in M_i$ is a dominant strategy at $\theta_i \in \Theta_i$, if

$$v_i(g_d(m_{-i}, m_i), \theta_i) + g_{t,i}(m_{-i}, m_i) \geq v_i(g_d(m_{-i}, \hat{m}_i), \theta_i) + g_{t,i}(m_{-i}, \hat{m}_i) \qquad (4)$$

for all m_{-i} and $\hat{m}_i$.

A dominant strategy has the strong property that it is optimal for a player no matter what the other players do. When dominant strategies exist, they provide compelling predictions for strategies that players should employ. However, the strong properties required of dominant strategies limits the set of situations where they exist.

A social choice function $f = (d, t)$ is implemented in dominant strategies by the mechanism (M, g) if there exist functions $m_i: \Theta_i \to M_i$ such that $m_i(\theta_i)$ is a dominant strategy for each i and $\theta_i \in \Theta_i$ and $g(m(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

3.12 Direct Mechanisms and the Revelation Principle

Note that a social choice function $f = (d, t)$ can be viewed as a mechanism, where $M_i = \Theta_i$

and $g = f$. This is referred to as a direct mechanism.

A direct mechanism (or social choice function) $f = (d, t)$ is dominant strategy incentive compatible if θ_i is a dominant strategy at θ_i for each i and $\theta_i \in \Theta_i$. A social choice function is also said to be strategy proof if it is dominant strategy incentive compatible.

The usefulness of the class of direct mechanisms as a theoretical tool in mechanism design is a result of the well-known, simple, and yet powerful revelation principle.

The Revelation Principle for Dominant Strategies

If a mechanism (M, g) implements a social choice function $f = (d, t)$ in dominant strategies, then the direct mechanism f is dominant strategy incentive compatible.

The Revelation Principle follows directly from noting that $f(\theta) = g(m(\theta))$ for each θ. The powerful implication of the revelation principle is that if we wish to find out the social choice functions can implemented in dominant strategies, we can restrict our attention to the set of direct mechanisms.

3.13 The Gibbard-Satterthwaite Theorem

Given that the specification of the space of decisions D can be quite general, it can keep track of all the aspects of a decision that are salient to a society. Thus, the transfer functions t are an extra that may be needed to provide correct incentives, but might best be avoided if possible. So, we start by exploring the set of decisions that can be implemented in dominant strategies without having to resort to transfers (beyond any that society already wished to specify inside the decisions), or in other words with t set to 0. A decision rule d is dominant strategy incentive compatible (or strategy-proof) if the social choice function $f = (d, t^0)$ is dominant strategy incentive compatible, where t^0 is the transfer function that is identically 0.

A decision rule d is dictatorial if there exists i such that $d(\theta) \in \text{argmax}_{d \in R_d} v_i(d, \theta_i)$ for all, where $R_d = \{d \in D \mid \exists \theta \in \Theta : d = d(\theta)\}$ is the range of d.

Theorem 1 Suppose that D is finite and type spaces include all possible strict orderings over D. A decision rule with at least three elements in its range is dominant strategy incentive compatible (strategy-proof) if and only if it is dictatorial.

(For any ordering $h : D \rightarrow \{1, ..., \#D\}$ (where h is onto) of elements of D and $i \in N$ there exists a type $\theta_i \in \Theta_i$ such that $v_i(d, \theta_i) < v_i(d', \theta_i)$ when $h(d) < h(d')$.)

The condition that type spaces allow for all possible strict orderings over D, is quite natural in situations such as when the set of decisions is a set of candidates, one of whom is to be chosen to represent or govern the society. But this condition may not be appropriate in settings where the decisions include some allocation of private goods and individuals each prefer to have more of the private good, as in an auction setting. The Gibbard-Satterthwaite theorem has quite negative implications for the hopes of implementing non-trivial decision rules in dominant strategies in a general set of environments. It implies that transfer functions will be needed for dominant strategy implementation of non-dictatorial decision rules in some settings. Before discussing the role of transfer functions, let us point out some prominent settings where the preferences do not satisfy the richness of types assumption of the Gibbard-Satterthwaite theorem and there exist non-dictatorial strategy-proof social choice functions that do not rely on transfer functions.

-
-
-

Bibliography

Corchon L. (1996). *The Theory of Implementation of Socially Optimal Decisions in Economics*, 165 pp. New York: St Martin's Press. [A text that covers dominant strategy mechanism design and implementation theory.].

Fudenberg D. and Tirole J. (1993). *Game Theory*. Cambridge, MA: MIT Press. [A text with a chapter devoted to Bayesian mechanism design.].

Jackson M.O. (2001). *A Crash Course in Implementation Theory, Social Choice and Welfare, Vol.18, No.4, pp655-708*. [An overview of implementation theory with a comprehensive and up- to-date bibliography.].

Mas-Colell A., Whinston M.D., and Green J.R. (1995). *Microeconomic Theory*, 981 pp. New York: Oxford University Press. [A text with a chapter devoted to mechanism design.].

Moore J. (1992). Implementation in Environments with Complete Information. (Ed. J.J. Laffont), *Advances in Economic Theory*. (Proceedings of the 6th Congress of the Econometric Society, in Barcelona, Spain, 1990). Cambridge University Press. [A survey of implementation theory including dominant strategy mechanism design.].

Moulin H. (1991), *Axioms of Cooperative Decision Making*, 332 pp. Cambridge: Cambridge University Press. [A text with chapters including aspects of dominant strategy mechanism design as well as results on public goods mechanisms.].

Palfrey T.R., and Srivastava S. (1993). *Bayesian Implementation*, 104 pp. Switzerland: Harwood Academic Publishers. [A monograph that surveys Bayesian implementation.].

Biographical Sketch

Matthew O. Jackson is the Edie and Lew Wasserman Professor of Economics, California Institute of Technology. He joined the faculty of the California Institute of Technology in 1997. Prior to that, he was on the faculty at Northwestern University from 1988 to 1997, and held the IBM Chair in Competitive and Regulatory Policy. He received a Ph.D. from Stanford University in 1988 and a B.A. from Princeton University in 1984. Jackson's current research includes implementation theory, the formation of social and economic networks, social choice theory, the design of markets, and game theory. Jackson's previous research includes studies of the design of futures contracts and the role of information in the formation of securities prices. Jackson is a fellow of the econometric society and a member of the editorial boards of *Econometrica*, *Games and Economic Behavior*, the *Journal of Economic Theory*, the *Journal of Public Economic Theory*, *Mathematical Social Sciences*, the *Review of Economic Design*, and *Social Choice and Welfare*.